Towards Active Event Recognition*

Dimitri Ognibene, Yiannis Demiris
Personal Robotics Lab,Imperial College London, UK
{d.ognibene,y.demiris}@imperial.ac.uk

Abstract

Directing robot attention to recognise activities and to anticipate events like goal-directed actions is a crucial skill for human-robot interaction. Unfortunately, issues like intrinsic time constraints, the spatially distributed nature of the entailed information sources, and the existence of a multitude of unobservable states affecting the system, like latent intentions, have long rendered achievement of such skills a rather elusive goal. The problem tests the limits of current attention control systems. It requires an integrated solution for tracking, exploration and recognition, which traditionally have been seen as separate problems in active vision. We propose a probabilistic generative framework based on information gain maximisation and a mixture of Kalman Filters that uses predictions in both recognition and attention-control. This framework can efficiently use the observations of one element in a dynamic environment to provide information on other elements, and consequently enables guided exploration. Interestingly, the sensors control policy, directly derived from first principles, represents the intuitive trade-off between finding the most discriminative clues and maintaining overall awareness. Experiments on a simulated humanoid robot observing a human executing goal-oriented actions demonstrated improvement on recognition time and precision over baseline systems.

1 Introduction

Anticipating the evolution of the external environment, which may comprise other agents, is essential to prepare and produce effective action responses [Pezzulo and Ognibene, 2012; Pezzulo, 2008; Balkenius and Johansson, 2007]. In this context an autonomous agent has to face three main difficulties: 1) acquire a predictive model of the environment; 2) compute predictions in a limited time [Zilberstein and Russell, 1996]; 3) control its sensors to perceive the current state of the environment and predict its evolution.

Figure 1: Experimental setup: iCub looking at target objects and arm movements (bottom right). The top-left image shows the iCubs gaze following the hand. In the top-right, hand movements of the human are anticipated by the iCub gaze. Finally, the bottom-left image shows gaze focused on the target object before the hand reaches it. These three images are grabbed from the iCub camera during an interaction trial.

Active Perception for Anticipation. This paper focuses specifically on the third problem, which can be named 'Active Perception for Anticipation' (APA). It consists in finding effective sensor control strategies to gather the information necessary to feed the available predictive models. The APA problem is proposed as a new component of the active vision paradigm [Aloimonos *et al.*, 1988; Ballard, 1991; Bajcsy, 1988; Suzuki and Floreano, 2008], which complements and integrates other components like active monitoring during action execution [Sridharan *et al.*, 2010], object detection [Vijayanarasimhan and Kapoor, 2010; de Croon and Postma, 2007; Vogel and de Freitas, 2008], tracking [Sommerlade and Reid, 2008; Gould *et al.*, 2007] and active object recognition [Paletta *et al.*, 2005; Denzler and Brown, 2002]. These components exploit task-specific knowledge to improve perception performance and reduce computational cost, but active vision has also shown to play an important role in improving learning performance [Andreopoulos and Tsotsos, 2013; Walther *et al.*, 2005] even when the task is not known [Og-

*This research was funded by EFAA FP7-ICT-Project, Grant Agreement no: 270490. The authors want to thank Yan Wu, Miguel Sarabia, Kyuhwa Lee, Eris Chinellato, Helgi P. Helgason, Margarita Kotti, Harold Soh, Sotirios Chatzis and Giovanni Pezzulo for their suggestions and IJCAI reviewers for the helpful comments.

nibene *et al.*, 2010]. An example, which can help understanding the APA problem, is the prediction of the motion of one element of interest (EI), say an asteroid, in an environment composed of multiple elements (e.g., asteroid field) with known dynamics, i.e. ideal motion and bouncing, when the state of some elements is not precisely known and the sensors receptive field is limited but controllable. A short time prediction of the trajectory of the EI can easily be produced by just tracking it and passively detecting other elements, which can give rise to an interaction and change the motion of the EI (e.g., an impact) when they enter into the field of view.

In most conditions longer term predictions are desired to produce more effective responses. The reliability and temporal reach of the prediction (e.g., 1 minute without impacts at 99% probability) depends on the sensor control strategy. The current motion of the EI is a *'dynamic context'* which can be used to actively find other elements who could produce an interaction. The agent can predictively direct its sensors along the trajectory of the EI and successively explore the areas close to the trajectory. The system should also come back to track the EI regularly, otherwise it may lose the EI (e.g., due to an impact with a very fast undetected element). Thus an effective attentional strategy is necessary to produce longer and more reliable predictions.

When the interactions can take place over long ranges, like when attraction and repulsions forces are present, the motion of each element is part of the 'dynamic context' giving information related to the state and presence of the other relevant elements (e.g., if an asteroid is accelerating the source of attraction can be found in the direction of the acceleration). Simultaneously, having more information on the elements in the environment can allow for more precise and longer predictions (e.g., predicting not only the impacts but the non-linear trajectories too, which in turn helps predicting impacts). This increases the importance of an active sensor strategy that explores some candidate areas for the presence of long-range interacting elements.

Active Event Recognition. In this paper we are interested in an even more complex condition, where the actual interaction between elements (e.g., repulsion, attraction or just bouncing) depends on some hidden states of the elements themselves (e.g., electric charge). In this case accurate estimations of motion and position of some elements can contribute at unveiling the value of the hidden state of other elements. In this condition, an efficient active sensor-control policy has the additional role of selecting when and what elements to track to permit the necessary estimations to recognizes the most informative elements (e.g., charged ones). We define Active Event Recognition (AER) as a sensor control strategy aimed at unveiling class of the event, which is the value of the hidden states determining the successive evolution of the environment (see figure 2). AER is thus an essential part of the APA problem.

While AER, like all the previous conditions, shows the importance of an active control of sensors for effective predictions in dynamic environments, AER is of particular interest because it can be directly mapped to broad set of conditions comprising goal-directed actions executed by other agents, e.g. a hand reaching a cup (see figure 1), two person going one toward the other, a car avoiding an obstacle, or a prey escaping from a predator. Anticipating such behaviourally-relevant[1] events still poses particularly demanding challenges in terms of the timely detection of relevant elements in an event and the recognition of the discriminative dynamics (e.g., motion of the hand). When the state of an element (e.g., prey) is uncertain, the 'dynamic context' (presence and dynamics of other elements e.g., a running predator) can provide valuable information for prediction. In general, to anticipate these 'non-local' events, and to produce an AER, the observer must couple its sensing behaviours with the independent dynamics of the environment which is yet unknown to the observer. The solution to this chicken-and-egg problem demands for a principled integration of tracking, exploration, search and recognition capabilities. However, these are perceived as separate problems in the active vision literature.

A principled integration requires the selection of the next sensor configuration (e.g., stop tracking the demonstrator and look for an graspable object) by merging, evaluating and exploiting different sources of information (e.g., noisy observations of hand movements and target positions). Solving this online is computationally complex and knowledge demanding. In particular the evaluation of the informative contribution of the dynamic context may require the prediction of the distribution of the expected trajectories[2].

Related Works. Some previous attempts of principled visual attention system, such as [Navalpakkam and Itti, 2006] which uses only local visual features, cannot direct attention to objects which are outside the visual field. Others, like Sprague and Ballard [2007], formalise the role of attention to subserve action execution, using independent models for the elements in the environment. Thus they cannot predict the elements interactions, like others' actions, and employ it for action selection or attention allocation. Attention systems, like [de Croon and Postma, 2007; Paletta *et al.*, 2005; Denzler and Brown, 2002], utilise the contextual information connected to low-level visual features and suffer from limited generalisation capabilities and reduced applicability to dynamic environments.

Proposed Approach. To achieve the efficient spatial-temporal coupling between the agent's sensors and the environment, we propose a probabilistic generative framework based on a *mixture of Kalman Filters* (KFs). It exploits several KF properties, like fast analytical update and computation of entropy, to reduce the computational complexity and evaluate the dynamic context. The sensor-control policy is directly derived from the principle of maximisation of expected information gain. It explicitly attempts at reducing the uncertainty on the event which is currently taking place.

[1] Anticipation and prediction of others' goals is at the basis of most human-human and human-robot interactions [Demiris, 2007].

[2] In the case of agent actions, this corresponds to the online solution of the intention recognition problem [Baker *et al.*, 2009; Ramirez and Geffner, 2011; Demiris, 2007]. This is a hard problem which can be further complicated by partial observability and exacerbated by the uncertainty of the environment.

Swift evaluation of the dynamic context can be achieved by making two assumptions: Firstly, events can be represented as linear dynamic systems[3]. Consequently, the state of the dynamic system implicitly characterises the expected kinematics. Secondly, a limited number of elements participate in each event. The expensive multidimensional optimisation for selecting the next sensor configuration can be approximated by choosing the configuration that maximises the expected information gain for the event from a set of candidates. Effective candidates can be built by reusing the predictions of the KF. This also allows the system to focus its sensors to positions outside of their current field of view and to select between visually similar elements, thus overcoming the limits of some other approaches like [Navalpakkam and Itti, 2006].

We apply this framework to the problem of directing the attention of a humanoid robot iCub to perceive a goal directed action, an archetype of non-local event. To furnish the necessary models of events we follow [Demiris and Khadhouri, 2008] and the simulation theory of action perception reusing the trajectory planner knowledge of the robot [Gori *et al.*, 2012]. The transition function of each KF of the mixture is an instance of the same stable linear dynamic system that is used in the planner but has a different attractor corresponding to a different target. [Demiris and Khadhouri, 2008] recognize actions by running in parallel the different models corresponding to the various action hypotheses. It controls attention through the direct competition between the models to access the information they need. It results in the selection of the elements which are necessary to control the currently winning action[4]. If the predictions of the different models for the selected elements are similar then the observations may not be useful for the recognition of the event. Instead our system directs attention *to directly discriminate between the different events* using information gain maximisation.

Using a robot simulator and synthetic and recorded human goal-directed actions, we compare our framework with the approach we proposed in [Ognibene *et al.*, 2012]. The latter extends [Demiris and Khadhouri, 2008] by integrating its predictive models with separate KF for each element in the environment.

2 Active Event Recognition

In this section the AER is defined and a solution based on a mixture of KF using Information Gain (**AERIG**) is described.

Problem definition. The graphical model in figure 2 displays the formulation of the problem. The discrete hidden stochastic variable V represents the class of the event which is taking place, characterised by a different dynamic of the environment that the agent must predict and recognise. The environment is composed of a fixed set of elements $I = \{i = 1 \dots n\}$ and thus its state $\mathcal{X}^k$ at time k is composed of the states X_i^k of the different elements. For each value of V the evolution of $\mathcal{X}^k$ is determined by a different dynamic system with different independence conditions between the elements. At each time step the agent receives for each element i an observation o_i^k which depends on the current configuration of the sensors θ^k. The states and observations are continuous variables.

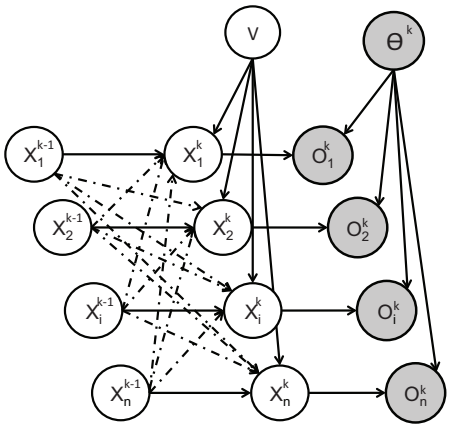

Figure 2: Perception of a non-local event. The discrete variable V represents the class of the event. The variables X_i are the state variables of various elements in the environment. Their next state is determined by their previous state and by V. The dashed connections indicate that the connection are sparse and that an element X_i is affected by an element X_j with $i \neq j$ only for a limited set of values of V. Each element i provides a different observation O_i^k which depends on the current state of the element and on the sensor configuration Θ.

At every time step the goal of the system is to select the configuration $\hat{\theta}^k$ that will minimise the expected uncertainty over V (quantified by entropy H):

$$\hat{\theta}^k = \underset{\theta^k}{\operatorname{argmin}} \int_{\mathbf{O}} p(\mathbf{o}^k|\theta^k) H(V|\mathbf{o}^k, \theta^k) \mathrm{d}\mathbf{o}^k \tag{1}$$

Proposed solution. For the recognition of the event and for the selection of the sensors configuration it is necessary to compute the posterior $P(v|\mathbf{o}^k; \theta^k)$. Given a prior distribution $P(v, \mathbf{x}_{1:N}^k) = P(\mathbf{x}_{1:N}^k|v)P(v)$ and the independence of the observed event from the sensor configuration $P(v|\theta) = P(v)$, the update expression of the posterior $P(v|\mathbf{o}^{k+1}\theta^{k+1})$ can be derived through the use of the Bayes rule:

$$P(v|\mathbf{o}^{k+1}, \theta^{k+1}) = \frac{P(\mathbf{o}^{k+1}|v, \theta^{k+1})P(v)}{P(\mathbf{o}^{k+1}|\theta^{k+1})} \tag{2}$$

The computation of eq.1 and eq.2 in the general case can pose severe computational complexities. The solution proposed is based on the assumption that, once v is fixed, the dynamics is linear and the probability distributions are normal. This enables the use of a mixture of KF with a distinct KF for each value of v. Denoting with $\bar{\mathbf{o}}_{v,\theta^{k+1}}^{k+1}$ the mean expected

[3]While the coupling between an agent and its environment as a dynamic system has longly been studied in various disciplines like artificial life and robotics [Nolfi and Floreano, 2000; Beer, 1995], it did not receive attention in the field of action recognition, where it can deliver several computational advantages over normative methods [Baker *et al.*, 2009] and is more parsimonious in terms of knowledge allowing direct use without extensive environment based training, necessary for models like [Bruce and Gordon, 2004]

[4]Also this attention system results in predictive gaze saccades toward the action target. The use of predictive models to control attention during action perception is supported by some experimental results like [Flanagan and Johansson, 2003].

observation and with $\mathbf{S}^{k+1}_{v,\theta^{k+1}}$ its covariance matrix, both of which are conditioned on v and θ and computed during the KF update, the following can be derived:

$$\hat{\theta}^{k+1} = \underset{\theta^{k+1}}{\operatorname{argmin}} \sum_v P(v) \Big(\frac{1}{2} \ln |\mathbf{S}^{k+1}_{v,\theta^{k+1}}| + \int_{\mathbf{O}} \mathcal{N}(\mathbf{o}; \bar{\mathbf{o}}^{k+1}_{v,\theta^{k+1}}, \mathbf{S}^{k+1}_{v,\theta^{k+1}}) \ln(P(\mathbf{o}|\theta^{k+1})) \mathrm{d}\mathbf{o} \Big) \tag{3}$$

Where $|S|$ is the determinant of a matrix S. The first order Taylor expansion of $P(\mathbf{o}|\theta)$ at point $\bar{\mathbf{o}}^{k+1}_v$ results in:

$$\hat{\theta}^{k+1} \approx \underset{\theta}{\operatorname{argmin}} \sum_v P(v) \left[\frac{1}{2} \ln |\mathbf{S}^{k+1}_{v,\theta^{k+1}}| + \ln \sum_{v'}^{V} \left(P\left(v'\right) \mathcal{N}(\bar{\mathbf{o}}_{v,\theta}; \bar{\mathbf{o}}^{k+1}_{v',\theta^{k+1}}, \mathbf{S}^{k+1}_{v',\theta^{k+1}}) \right) \right] \tag{4}$$

The first term of eq. 4 is the average of the expected entropy of the observations for each model v on the new observation point. The second term is a measure of how much, in the new sensor configuration, the observations predicted by a model will differ from those predicted by other models. Thus, the proposed formulation integrates a trade-off between discriminating the event hypotheses and maintaining their perception quality.

We will now use the assumption that to each event (for each value of v) involved only a limited subset of elements of the environment, the set $I_v \subset I$. The dynamics of these elements will be coupled, formally their transition probabilities depends on the state of the other elements of I_v, that is $P(\mathbf{x}^k_i|\mathbf{x}^{k-1}_{1:N_e}, v) \equiv P(\mathbf{x}^k_i|\mathbf{x}^{k-1}_{I_v}, v), i \in I_v$. The dynamic of those elements which do not participate in the event will be independent of each other, $P(\mathbf{x}^k_i|\mathbf{x}^{k-1}_{1:N}, v) \equiv P(\mathbf{x}^k_i|\mathbf{x}^{k-1}_i), i \notin I_v$. This allows decomposition of each KF corresponding to a value of v in a set KFs of lower dimensions. One of the KF will model the coupled dynamic of the elements in I_v while the dynamic of each element $i \notin I_v$ will be modelled by an independent KF. This decomposition delivers two advantages: 1) The computational cost of updating a KF is cubic in the dimension of the state space. Thus, updating a set a lower dimension KFs is more efficient than updating a single high dimensional one; 2) The KFs for the elements which do not belong to I_v are independent of v, which permits to reuse the results of the KF update.

The KF decomposition can be utilised for efficient computation of eq. 4. The first term can be computed as the product of the determinant of the covariance matrix $\tilde{\mathbf{S}}^{k+1}_{v,\theta}$ of the observation distribution predicted by the KF of I_v and those $\dot{\mathbf{S}}^{k+1}_{i,\theta}$ from the KF of the other elements.

$$\ln |\mathbf{S}^{k+1}_{v,\theta}| = \ln |\tilde{\mathbf{S}}^{k+1}_{v,\theta}| + \sum_i^{I-I_v} \ln |\dot{\mathbf{S}}^{k+1}_{i,\theta}| \tag{5}$$

The second term of eq. 4 can also be decomposed in a similar way leading to reduced computational complexity.

Solving directly eq. 4, even after decomposition, can still be complex. Instead, a finite set $\hat{\Theta}$ of candidate sensor-configurations is defined and the one with the highest expected information gain is selected. In this work, the elements of $\hat{\Theta}$ are the configurations which will minimise the noise on any of the elements according to any of the possible hypotheses. E.g., $\hat{\Theta}$ can be the set of different camera configurations focusing on the predicted positions of the different objects according to the different possible events.

3 Experimental Evaluation: Active action recognition

The framework proposed in the previous section, AERIG, has been applied to control the gaze of an iCub humanoid robot that has to anticipate the target of a reaching action performed by a human partner (see figure 1). In the tests the robot showed human-like attentional behaviour while observing human actions, switching its attention from tracking the hand to the anticipated target (see videos at **`http://www.imperial.ac.uk/personalrobotics`**).

The most important contribution of the approach is the predictive recognition of the event and the temporal advantage for action preparation it permits. However, due to human reaction delays, it is complex to assess the temporal performance of the approach with a real robot. In fact, during the tests, unlike during normal operations, it is important to compare synchronously the timings of performer's intention and action with the evolution of the robot estimations. Instead simulation results are reported with different versions of the sensor model: with a fixed camera (**NO V**isual **A**ttention **C**ontroller, **NOVAC**) randomly placed near one of the elements, an ideal camera with instantaneous and precise gazing movements (**IDEALVAC**), and with the simulated camera controller reproducing the speed and trajectories of the real iCub camera controller [Pattacini, 2010] (**REALVAC**). The robot's perception of the position of an element (object or of the other agent hand) is considered affected by gaussian noise whose variance is a linear function of the distance between the gaze focus point and the element ($\sigma = 0.1 + 0.2 \times d$) [5]. This observation model is a simplification of the human vision which has greatly higher resolution in the centre [Findlay and Gilchrist, 2003]. The system was tested on both synthetic data, generated by the same control model used by the robot, and on natural human reaching actions recorded with the Kinect. AERIG is compared to a heuristic attention controller proposed in [Ognibene *et al.*, 2012] (**HEURISTIC**). The heuristic balances element position uncertainty with event uncertainty by selecting the next target to gaze as the element with the highest product between its position entropy and the confidence of the event in which it takes part[6]. In this task, the environment (I) is composed by the hand of the human performer (h) and by a fixed number N_e of graspable objects

[5] An object specific factor can be applied to reflect its intrinsic difficulty in recognition and localisation. This observation noise model is correlated with the state, thus the optimality properties of Kalman filter cannot be guaranteed.

[6] The effector position is estimated using a KF and the same motion model used by AERIG ([Gori *et al.*, 2012]). The target position estimation needed by the motion model is estimated by an independent KF which model assumes the target to be still, this does not allow to use the motion of the effector to correct the target position. The confidence c is updated at every time step using the prediction error e^k_h on the hand position $c^{k+1} = c^k + (1 + e^{k+1}_h)^{-1}$

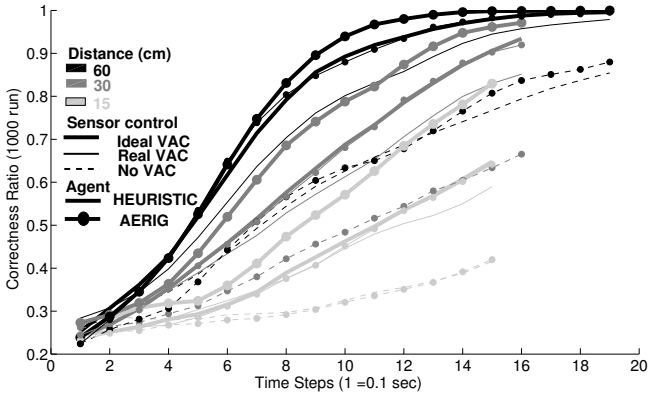

Figure 3: Evolution of the average success rate on simulated data with 4 objects.

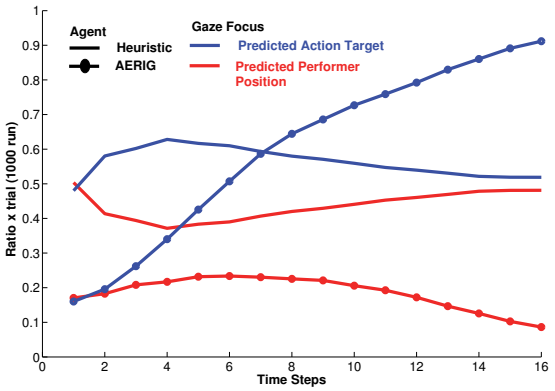

Figure 4: Evolution of the average gaze focus during a trial with REALVAC and 30 cm of distance between the targets. It shows only the saccades directed towards the elements of the currently most probable action.

$(i = 1 \dots N_e)$. Each event v corresponds to a hand movement to reach a different object in the environment (thus the domain of V is $\{1 \dots N_e\}$). The state X_i^k of an object i represents its position, while the state X_h^k of the hand consists of both speed and position. The set of coupled elements I_v is composed by the hand and by the the object with index $i \equiv v$. The KF has the same transition matrix which is used to control a reaching action execution, thus complying with the simulation theory of action perception [Demiris, 2007; Dindo *et al.*, 2011]. The equation employed by a given model v to compute the next effector position $\mathbf{x}_{p,v}^{k+1}$, when the target is at position $\mathbf{x}_v^k$:

$$\mathbf{x}_{p,v}^{k+1} = \mathbf{x}_{h,v}^k + \tau \left\{ \dot{\mathbf{x}}_{h,v}^k + \tau \left[K(\mathbf{x_v} - \mathbf{x}_{h,v}^k) - D\dot{\mathbf{x}}_{h,v}^k \right] \right\} \quad (6)$$

K and D are typical constants of a PD controller while τ is the related time integration parameter. They are set to the same default values used for action generation $K = 1.5$, $D = 3$ and $\tau = 0.16$, τ can be modulated to model different motion speeds. The Kalman filter associated with each single element assume that the element is still (thus the transition matrix is the identity) with its position affected only by zero-mean Gaussian noise with variance 0.0025. The actual system is open source and available online at **`http://www.imperial.ac.uk/personalrobotics`**.

3.1 Results

Figure 3 reports the evolution of the"success rate"; how many times the currently estimated event is actually the correct one, with simulated effector trajectories in an environment with 4 objects at different distances from each other. The average duration of trial depends on the distance between the objects. In each condition 1000 runs were executed. The figure shows significant improvement (up to two times better) when AERIG is adopted, delivering more accurate and faster recognition in all conditions, specially when the task is more difficult. It also shows the crucial role of active sensor control. Similar results were observed with a varying number of objects, thus showing that AERIG can scale with the number of elements. On the computational side, on a conventional 2010 PC (2.8Ghz), AERIG takes 0.4ms for frame with 16 targets. Figure 4 displays the average evolution of the attentional behaviour of the AERIG and HEURISTIC agents with the real gaze controller. The graph shows that both the systems do not focus just on the effector but continuously alternates between it and the targets. Thus AERIG, which uses the effector motion information to correct the estimated position of the target, can more precisely direct its saccades towards the target and thus improve the overall prediction. The AERIG system explores the environment and does not focus only on the current most probable target and effector position. Successively it increases the saccades towards those elements and in particular the target. The HEURISTIC model instead focuses only on the elements of the most probable event. Figure 5 displays the results of AERIG when applied to human reaching actions recorded with the Kinect. The 3D positions of the right hand of the user skeleton, extracted using the OpenNI [Ope, 2010] and NITE [Pri, 2010] libraries, were recorded during 24 reaching actions towards 4 objects positioned on a circle of radius 15 cm (see figure 6). The actions, with an average duration of 1 sec, were performed naturally from different starting positions at an average distance of 40 cm from the circle centre. At each time step, the actual position perceived by the robot was affected by noise depending on the current gaze position in the same way as in the previous ex-

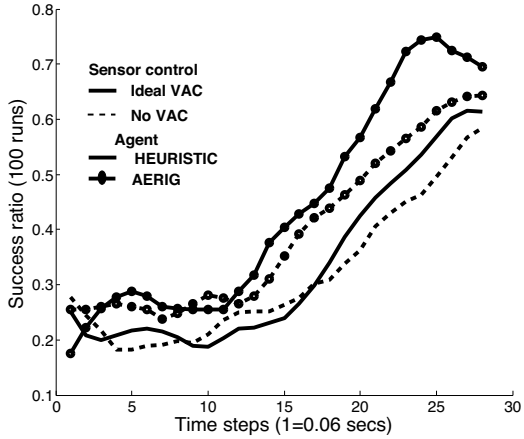

Figure 5: Evolution of the average success rate on recorded human action with 4 objects at a distance of 15 cm.

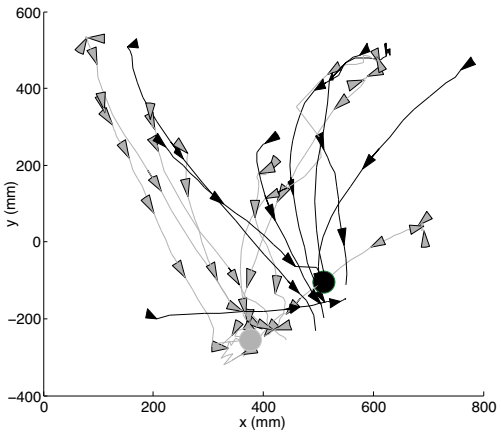

Figure 6: Samples of recorded reaching action trajectories (2d projection) towards two different objects (circles) .

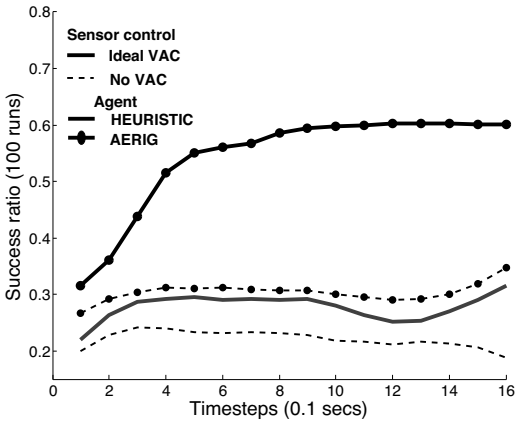

Figure 7: Evolution of the average success rate of the system with limited field of view.

periment. It is easy to see that the trajectories are not straight as predicted by the models used.

This data show that AERIG delivers improvement in event recognition even when its knowledge is noisy and does not match the real dynamic of the events. While both AERIG and the heuristic model use the same action model, and thus suffer from similar problems with the recorded data, AERIG is more robust because it is sensitive to uncertainty when selecting the actual event. We also hypothesise that the robustness of AERIG is partially due to the use of stable dynamic systems to model the interactions in the environment. The trajectories are also affected by noise which is specially evident near the target objects. This noise cannot be reduced by the active control of the camera. This helps explain the shape of the performance in Figure 5. Figure 7 displays the performance of the systems when the robot sensor has a limited field of view (radius of 40cm around the gaze focus) and reduced noise ($\sigma = 0.03 + 0.06 * d$), with four objects at a distance of 30 cm. This result, while preliminary, shows that only AERIG with active camera can provide useful events recognition[7]. This is due to the use of the effector motion information which is used to adjust the initial random guesses of the target position, which in the other conditions were based on the peripheral vision. This adjustment are successively used to confirm the presence of the target with direct gazes.

4 Conclusions

This paper introduces AERIG, a probabilistic framework for the active perception of multi-element dynamic events, like goal-oriented actions. It addresses the requirements of the APA problem, which consists of finding effective sensor control strategies to gather the information necessary to predict the evolution of the environment. AERIG actively directs sensors to the most discriminative configurations to uncover the actual dynamics of the environment. This leads to the spatio-temporal coupling of the sensors with the environment dynamics.

While actively recognising a complex dynamic context [Rothkopf *et al.*, 2007] has crucial behavioural importance, this problem has been not sufficiently addressed in the literature of attention. In fact, most of the attention models assume artificial laboratory conditions and offer limited explanations of task oriented attention in dynamic environments [Tatler *et al.*, 2011]. Thus, it would be intriguing to compare the predictions of the proposed framework with actual human behaviour in natural conditions. AERIG applied to the recognition of goal oriented actions represents a sound and robust implementation of the simulation theory of action perception [Demiris, 2007]. While some theories [Flanagan and Johansson, 2003] support the reuse of the same attention control schemas used during action execution, our approach focuses on the direct reduction of uncertainty during the perception of the action. In this task the tests showed substantial performance improvement over baseline approaches. Interestingly the framework allows a mathematical formulation of the intuitive trade-off between a conservative behaviour, which tries to observe all the events and elements, and a greedy one that is only interested in determining the real event which is taking place.

In this work, we tested our approach only on stable approximately linear systems with a limited number of elements and only used position information. In the future, we plan to extend the framework to use more complex models and integrating it with the overall agent control architecture.

[7]The policy used by the systems to manage missing observations was: a) if the element was predicted to be observed then generate a random observation outside of the field of view and reduce the probability of the event to the minimum event probability, b) otherwise generate a noisy observation of the element centred on the predicted position.

References

[Aloimonos *et al.*, 1988] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, january 1988.

[Andreopoulos and Tsotsos, 2013] A. Andreopoulos and J. K. Tsotsos. A computational learning theory of active object recognition under uncertainty. *Int. J. Comp. Vis.*, 101(1):95–142, 2013.

[Bajcsy, 1988] R. Bajcsy. Active perception. *Proceedings of the IEEE, Special issue on Computer Vision*, 76(8):966–1005, 1988.

[Baker *et al.*, 2009] C. L Baker, R. Saxe, and J. B Tenenbaum. Action understanding as inverse planning. *Cognition*, 2009.

[Balkenius and Johansson, 2007] C. Balkenius and B. Johansson. Anticipatory models in gaze control: a developmental model. *Cognitive processing*, 8(3):167–174, 2007.

[Ballard, 1991] D.H. Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.

[Beer, 1995] R. D. Beer. A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72:173–215, 1995.

[Bruce and Gordon, 2004] A. Bruce and G. Gordon. Better motion prediction for people-tracking. In *Proc. of the 2004 IEEE International Conference on Robotics and Automation*, New Orleans, LA, USA, May 2004.

[de Croon and Postma, 2007] G. de Croon and E.O. Postma. Sensory-motor coordination in object detection. In *Proceedings of the IEEE Symposium on Artificial Life, ALIFE'07*, pages 147–154, 2007.

[Demiris and Khadhouri, 2008] Y. Demiris and B. Khadhouri. Content-based control of goal-directed attention during human action perception. *Interaction Studies*, 9(2):353–376, 2008.

[Demiris, 2007] Y. Demiris. Prediction of intent in robotics and multi-agent systems. *Cognitive Processing*, 8(3):151–158, 2007.

[Denzler and Brown, 2002] J. Denzler and C.M. Brown. Information Theoretic Sensor Data Selection for Active Object Recognition and State Estimation. *Transactions in Pattern Analysis and Machine Intelligence*, 24(2):145–157, 2002.

[Dindo *et al.*, 2011] H. Dindo, D. Zambuto, and G. Pezzulo. Motor simulation via coupled internal models using sequential monte carlo. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, Barcelona, Spain, July 2011.

[Findlay and Gilchrist, 2003] J.M. Findlay and I.D. Gilchrist. *Active vision: The psychology of looking and seeing.* Oxford University Press, 2003.

[Flanagan and Johansson, 2003] J Randall Flanagan and Roland S Johansson. Action plans used in action observation. *Nature*, 424(6950):769–771, 2003.

[Gori *et al.*, 2012] I. Gori, U. Pattacini, F. Nori, G. Metta, and G. Sandini. Dforc: a real-time method for reaching, tracking and obstacle avoidance in humanoid robots. In *IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS2012),*, Osaka, Japan, November 29 - December 1. 2012.

[Gould *et al.*, 2007] S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. Bradski, P. Baumstarck, S. Chung, and A. Y. Ng. Peripheral-foveal vision for real-time object recognition and tracking in video. In *IJCAI07, Proceedings of*, 2007.

[Navalpakkam and Itti, 2006] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimal object detection. In *Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[Nolfi and Floreano, 2000] S. Nolfi and D. Floreano. *Evolutionary Robotics: The Biology,Intelligence,and Technology*. MIT Press, Cambridge, MA, USA, 2000.

[Ognibene *et al.*, 2010] D. Ognibene, G. Pezzulo, and G. Baldassarre. How can bottom-up information shape learning of top-down attention control skills? In *Proceedings of 9th International Conference on Development and Learning*, 2010.

[Ognibene *et al.*, 2012] D. Ognibene, E. Chinellato, M. Sarabia, and Y. Demiris. Towards contextual action recognition and target localization with active allocation of attention. In *Proc. of First Int. Conf.on Living Machines*, pages 192–203, 2012.

[Ope, 2010] OpenNI organization. *OpenNI User Guide*, November 2010. Last viewed 19-01-2011 11:32.

[Paletta *et al.*, 2005] L. Paletta, G. Fritz, and C. Seifert. Cascaded sequential attention for object recognition with informative local descriptors and q-learning of grouping strategies. In *Proc.of Computer Vision and Pattern Recognition (CVPR)*, 2005.

[Pattacini, 2010] U. Pattacini. *Modular Cartesian Controllers for Humanoid Robots: Design and Implementation on the iCub*. PhD thesis, RBCS, Istituto Italiano di Tecnologia, Genoa., 2010.

[Pezzulo and Ognibene, 2012] G Pezzulo and D Ognibene. Proactive action preparation: Seeing action preparation as a continuous and proactive process. *Motor control*, 16(3):386–424., 2012.

[Pezzulo, 2008] G. Pezzulo. Coordinating with the future: the anticipatory nature of representation. *Minds and Machines*, 18(2):179–225, 2008.

[Pri, 2010] PrimeSense Inc. *Prime Sensor NITE 1.3 Algorithms notes*, 2010. Last viewed 19-01-2011 15:34.

[Ramirez and Geffner, 2011] M. Ramirez and H. Geffner. Goal recognition over pomdps: Inferring the intention of a pomdp agent. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, Barcelona, Spain, 2011.

[Rothkopf *et al.*, 2007] C. A. Rothkopf, D. H. Ballard, and M. M. Hayhoe. Task and context determine where you look. *Journal of Vision*, 7(14):1610–1620, 2007.

[Sommerlade and Reid, 2008] R. Sommerlade and I Reid. Information theoretic active scene exploration. In *Proceedings of 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, USA, June 2008.

[Sprague *et al.*, 2007] N. Sprague, D. Ballard, and A. Robinson. Modeling embodied visual behaviors. *ACM Trans. Appl. Percept.*, 4(2):11, 2007.

[Sridharan *et al.*, 2010] M. Sridharan, J. Wyatt, and R. Dearden. Planning to see: A hierarchical approach to planning visual actions on a robot using pomdps. *Artificial Intelligence*, 174(11):704–725, 2010.

[Suzuki and Floreano, 2008] M. Suzuki and D. Floreano. Enactive robot vision. *Adaptive Behavior*, 16(2-3):122–128, 2008.

[Tatler *et al.*, 2011] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11(5):1–23, 2011.

[Vijayanarasimhan and Kapoor, 2010] S. Vijayanarasimhan and A. Kapoor. Visual recognition and detection under bounded computational resources. In *Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1006–1013, June 2010.

[Vogel and de Freitas, 2008] J. Vogel and N. de Freitas. Target-directed attention: Sequential decision-making for gaze planning. In *The International Conference of Robotics and Automation (ICRA)*, pages 2372–2379, 2008.

[Walther *et al.*, 2005] D. Walther, U. Rutishauser, C. Koch, and P. Perona. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding*, 100(1-2):41–63, October 2005.

[Zilberstein and Russell, 1996] S. Zilberstein and S. Russell. Optimal composition of real-time systems. *Artificial Intelligence*, 82(1-2):181–213, 1996.